A Usability Case Study Using TREC and ZPRISE

Laura L. Downey
National Institute of Standards and Technology (NIST)
Currently with: Vignette Corporation
901 S Mo Pac Expressway, Building 3
Austin, TX 78746-5776 U.S.A.
512-306-4467 phone
ldowney@vignette.com

Dawn M. Tice*
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Stop 8940
Gaithersburg, Maryland 20899-8940 U.S.A.
301-975-5055 phone
301-975-5287 FAX
dawn.tice@nist.gov

Working Title: NIST Usability Case Study

* For reprints: National Institute of Standards and Technology (NIST), Attn: Dawn Tice, 100 Bureau Drive, Stop 8940, Gaithersburg, MD 20899-8940. Also available from: http://www.itl.nist.gov/div894/894.02/works/papers.html.

Abstract

This paper examines the challenges involved in conducting an informal usability case study based on the introduction of a new information retrieval system to experienced users. We present a summary of activities performed during two iterations of usability testing and describe our analysis methodology. This methodology incorporates several grouping and prioritizing methods which provide one of the major contributions of the work. During the course of the case study, we learned some valuable lessons which were specific to the Text REtrieval Conference (TREC). The TREC-specific lessons learned led to recommendations for changes in the TREC topic development and assessment tasks. Results of the study include lessons learned about both the users and the testing techniques (Hoffman & Downey, 1997).

A Usability Case Study Using TREC and ZPRISE

Laura L. Downey[*], Dawn M. Tice[†]
National Institute of Standards and Technology, Gaithersburg, Maryland 20899, U.S.A.

Abstract

This paper examines the challenges involved in conducting an informal usability case study based on the introduction of a new information retrieval system to experienced users. We present a summary of activities performed during two iterations of usability testing and describe our analysis methodology. This methodology incorporates several grouping and prioritizing methods which provide one of the major contributions of the work. During the course of the case study, we learned some valuable lessons which were specific to the Text REtrieval Conference (TREC). The TREC-specific lessons learned led to recommendations for changes in the TREC topic development and assessment tasks. Results of the study include lessons learned about both the users and the testing techniques (Hoffman & Downey, 1997).

1 Introduction

The purpose of this study was three-fold: to gain experience in conducting usability testing on information retrieval systems, to specifically examine the usability of the new ZPRISE interface, and to identify problems our users were having with the assigned task, topic development for TREC-5 (NIST, 1998a).

TREC is the Text REtrieval Conference co-sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA). TREC's goal is to encourage research in information retrieval from large text applications by providing a large test collection, a common evaluation, and a forum for organizations to compare their results (NIST, 1998b). The test collections used for TREC-5 consisted of about 655,400 documents and 50 test questions (or topics). The task examined in this case study was the development of these topics for TREC-5.

In the first phase of the TREC process, NIST's users (called assessors) develop candidate topics for TREC by performing trial retrievals against a sample of the document set to be used for the TREC tasks. A subset of the candidate topics are selected for the TREC tasks. The TREC data sets contain thousands of documents and covers an extensive range of subject areas. Therefore, the users provide an invaluable service by developing topics that will retrieve "reasonable" numbers of relevant documents from these collections. Topics that are too narrow or too broad are not as effective for evaluation of systems – with too few relevant documents, evaluation metrics are unstable and with too many relevant documents, recall is underestimated.

---

[*] Currently with Vignette Corporation.
[†] To whom all correspondence should be addressed (E-mail: dawn.tice@nist.gov).

In the second phase, TREC participants run the selected test topics against the document set on their systems and send their results to NIST. TREC uses the pooling method (Sparck Jones & van Rijsbergen, 1975) to select a sample of the participants' results and provides these documents to the users for relevance judgments. The users' judgments serve as the basis for relevance on these topics, both for evaluation in a given TREC and as part of the permanent test collection (Voorhees & Harman, 1997).

The assessors used the ZPRISE interface during the topic development task and the relevance assessment task. The ZPRISE system, originally known as PRISE, developed at NIST in 1988 for the IRS as a prototype experimental statistical full-text searching system, demonstrated the usefulness of a ranked retrieval system in supporting free-form natural language queries against large document collections on minicomputers (Harman & Candela, 1990). Available in the public domain, PRISE has been incorporated into various NIST information retrieval projects and information retrieval projects outside of NIST. In 1995, PRISE became ZPRISE with the development of a client/server interface based on the ANSI/NISO Z39.50 standard (Library of Congress, 1998). Since the first release of ZPRISE, many functionality improvements have been made to the system. ZPRISE is currently being used by over 80 groups in a variety of ways: as a matrix to embed software, as a source of individual components, and to compare with other systems.

The introduction of the new ZPRISE interface provided an excellent opportunity to undertake a usability case study. The usability aspect of the old PRISE interface had never been tested. Since this was our first attempt in developing a client/server interface, we were certain there would be some usability issues to resolve with the new interface. Conducting this study would allow us to gain experience in usability testing while simultaneously examining the usability of the new interface and exploring the challenges the assessors faced with the topic development task.

2 Testing Methodology

The users for this study (the TREC assessors) are retired information analysts from the National Security Agency (NSA). Most of the users have been performing the topic development and relevance assessment tasks for NIST for four years within the TREC project.

For the usability study, we chose the TREC-5 topic development task. This allowed the actual TREC-5 topic development activity and the usability study to be conducted in parallel. Each user was instructed to compose topics on any subject of interest to them prior to the usability test. They were required to provide the following information for each topic: a short title, a short description of the topic, and a narrative that explained what would constitute a relevant document match.

Once the usability test began, the users searched a pre-selected database for their topics. During the search they marked documents relevant to the topic and also recorded the number of relevant documents found per topic. The users performed the searches using the NIST ZPRISE system which was installed on networked SUN Microsystems UNIX Workstations.

Prior to the actual usability test, the users answered questions on their topic development activities. This data was not part of the usability test but was gathered to support ongoing investigations into user search behavior.

Based on traditional usability practices, we chose a three-step process: a tutorial, observations and verbal feedback, and a satisfaction survey (Nielsen, 1993).

2.1 Usability Test 1

We conducted the first test with the following parameters:

- 11 users, all familiar with the topic development task
- 2 users per session, two 4-hour sessions per day

We began the test by giving a 1-hour tutorial to the users explaining the features of the new interface. One trainer demonstrated the new interface while the second trainer recorded the users' comments during the tutorial.

During the second part of the test, we observed the users for 50 minutes while they navigated the system performing their topic development task. When users had trouble, we encouraged them to problem-solve on their own or to consult system help or the written instructions. Users were given an additional 30 minutes to finish their topic development without observers in the room.

For the final portion of the usability test, we administered a user satisfaction survey (30-45 minutes).

2.2 Usability Test 2

Based on input from the first usability test, the interface was modified. A second usability test was conducted under the same conditions as Test 1, with four users (two of the original users and two new users).

2.3 Analysis

After conducting each iteration of the usability tests, we performed an analysis of the user responses applying several grouping and prioritizing methods. We grouped the user responses into three broad categories: critical incidents, in-scope, and out-of-scope factors. Critical incidents were defined as events the user was unable to solve without the observer intervening. An example of a critical incident occurred when several users were unable to start the topic development system without the observer’s assistance. Training and usability issues were categorized as in-scope factors. For example, training issues addressed problems in understanding the new graphical user interface (GUI) and usability issues often required a change in terminology used in the interface. Issues that were beyond our control, such as the time it takes for Openwindows to start, were regarded as out-of-scope factors. Next we estimated the time in staff-hours and cost associated with code changes. Factoring in the severity of the problems and the time/cost required to correct them resulted in the final decision model.

The challenge in analyzing all of the collected data was to organize it in order to identify the major system and interface issues. We gathered the following data per user:

- tutorial observations
- task performance observations including critical incident recordings
- user satisfaction survey data

3 Analysis of Usability Test 1

Figure 1 is a screenshot of the initial interface presented to the users during the first pass of usability testing. First, the users logged into the system by entering their username and password. The topic development system was initialized by typing "zclient" at the system prompt. The users' first step was to select a server (1). Then a database (2) was selected. Next the user entered the query (3) and clicked the "Perform Search" (4) button. The documents (5) that the system deemed relevant to the query were displayed in the "Document List" window. The user selected a document to be viewed by clicking on the title box of the desired document in the "Document List" window. The document (6) was displayed in the "Document" window and the user read the document to determine its relevancy to the query entered. The document was then judged (7) in the "Document List" window.

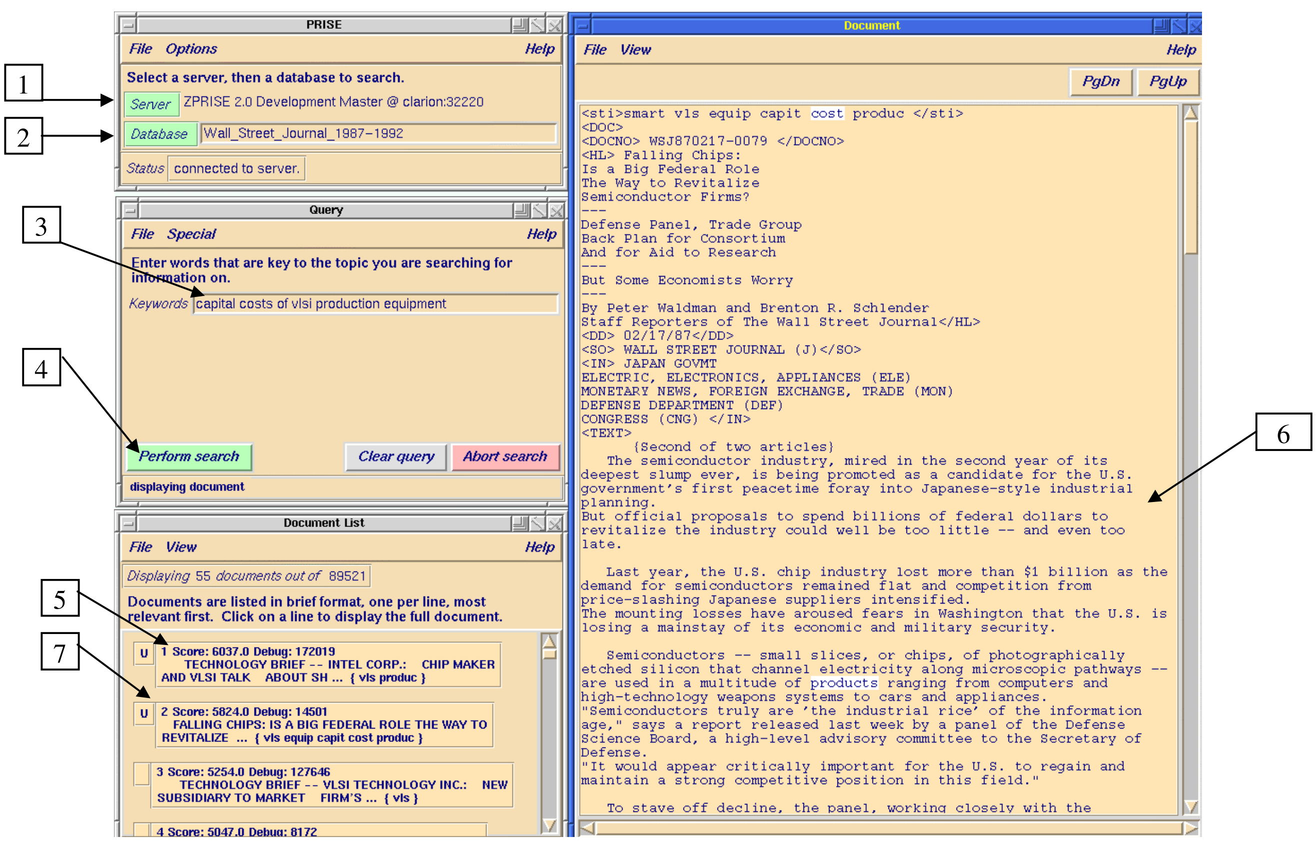

Figure 1 - Initial User Interface

Figure 2 shows the initial interface with the relevance feedback features. Relevance feedback is a statistical method used in information retrieval to automatically generate improved query statements. In ZPRISE, relevance feedback operates on the documents marked relevant and then retrieves documents that contain similar keywords. The user has the option of viewing only the documents judged relevant by selecting the proper choice under the "View" (1) button in the "Document List" window. This action displays the "Relevant Document List" (2) window. If the user is not satisfied with the documents that the system has retrieved, the system offers help in selecting query terms that may display documents that are more relevant to the query. By clicking on "Show Enhanced Query Terms" (3) in the "Relevant Document List" window, the "Enhanced Query Terms" (4) window is displayed. New query terms are added to the list by checking (5) the box adjacent to the term, selecting "add terms to query" (6) button and clicking on the "Perform Search" (7) button in the "Query" window. When all of the documents are judged and the user is satisfied with the judgments, then a new query is entered.

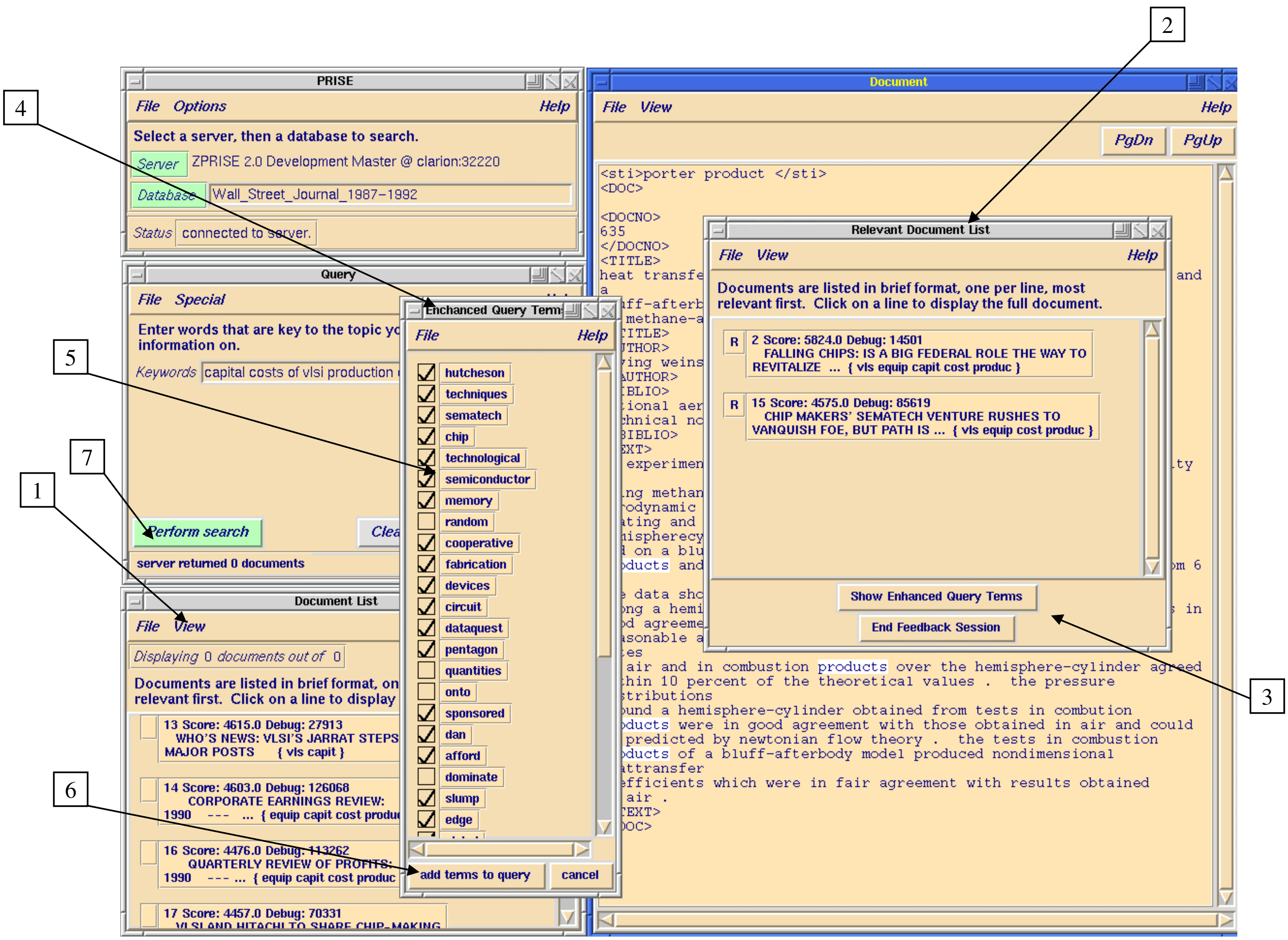

Figure 2 - Initial User Interface - Relevance Feedback

To start analysis of the test of this initial interface, we first combined the tutorial observations and the task performance observations by user. The combined observations were grouped by interface window, then by user to identify problem trends among users and common problems in specific parts of the interface.

We then consolidated similar problems and separated problems attributed to training issues to create a set of combined observations derived from the tutorial and task performance user responses. We also identified out-of-scope observations, such as problems related to the underlying windowing system and not to the interface itself.

Table 1 is a sample analysis of the observed and verbal feedback data gathered from our users. The data is displayed for each screen of the ZPRISE interface and users are identified by user number.

<table>
<tr><th colspan="2">Sample Analysis of Observed/Verbal Feedback Data by Screen/User</th></tr>
<tr><th colspan="2">PRISE Zclient Opening Window</th></tr>
<tr><th>Disposition</th><th>Observations/verbal feedback from users</th></tr>
<tr><td>
Address in training⇐
out-of-scope⇐
usability issue⇐
usability issue⇐
critical incident⇐</td><td>User 2
• trouble logging in
• did not want to wait for Openwindows startup
• tried selecting database first, then looked under options
• “what does default mean?”
• asked observer - “where do I get the assessor system”</td></tr>
<tr><td>
System problem⇐</td><td>User 4
• logged on without a problem, but “cd zclient” failed because the account was not set up</td></tr>
<tr><td>
Address in training⇐
(also critical incident)
usability issue⇐</td><td>User 6
• not totally clear how to start up the session (connect to server, etc.)
• did not easily locate server status messages</td></tr>
</table>

Table 1 - Sample Analysis of Observed/Verbal Feedback Data by Screen/User

NOTE: This data is abbreviated. The examples are shown to demonstrate organization and analysis, rather than to report the full observations/verbal feedback recorded.

At this point we incorporated the data from the satisfaction survey that was relevant to the identified usability issues in order to combine all the observed and perceived problems. We also created two other lists: a set of positive comments about the system and a list of users' suggestions for future enhancements. Table 2 represents a sample analysis of the data gathered during the satisfaction survey. The satisfaction survey identified questions pertaining to each screen of the ZPRISE interface. The users’ comments are identified by user number and the disposition of the comment is indicated.

Sample Analysis of Satisfaction Survey Data by Screen/Question/Users	
PRISE Zclient Opening Window	
Disposition	**satisfaction survey comments**
	Did each window/section make clear how it was to be used (i.e., what to do and in what order)?
Positive comment⇐	• Yes, easy (User 4)
Positive comment⇐	• Like date ranges displayed for the collections (User 3)
Positive comment⇐	• Same color used for the buttons was helpful (User 4)
Positive comment⇐	• Easy enough (Users 7 & 8)
Positive comment⇐	• Yes (User 5)
Positive comment⇐	• Logical steps - good (Users 2 & 9)
	If not, what was unclear?
Usability issue⇐	• Had trouble locating which collection was being displayed (Users 3 & 4)
Usability issue⇐	• Default in select server (User 10)
Usability issue⇐	• Also says "connected to server" - confusing (User 10)
	Are there any features you wish were available?
Positive comment⇐	• Likes that it shows collection and dates of collection (Users 3 & 4)
Wishlist⇐	• Would like to be able to put topic number in (Users 7 & 8)
	• Not really (User 2 & 9)
usability issue⇐	• it would be useful to include the document number with the document title instead of the document score (User 11)

Table 2 - Sample Analysis of Satisfaction Survey Data by Screen/Question/Users

NOTE: This data is abbreviated. The examples are shown to demonstrate organization and analysis, rather than to report the full satisfaction survey data.

As the final step in organizing the usability problem matrix, we categorized the items into several sub-groups such as problems relating to messages in the interface or information organization. We then assigned high, medium and low priorities to the problems as shown in Table 3. High priority was assigned to problems that directly affected the users' performance of the task. Problems that were deemed less serious, but still needed to be corrected, were assigned a medium priority. Code changes which might result in complicating the users' steps were given a low priority and a better solution to the problem was sought. With the development team, we proposed usability solutions and discussed the cost/benefit of each, resulting in a set of action items and estimates for changes to the interface. Table 3 depicts a sampling of the usability matrix used to categorize the problems the users identified.

Sample Usability Matrix by Problem Category				
Group	**Issue**	**priority**	**Time**	**action**
Ambiguous terminology	Functional overlap of “clear query” and “abort search”	M	Nominal	Change button label “abort search” to “abort search, clear results”
Fragmented functionality	would like to mark document relevant from document window and not from document list window	H	2 days	Add relevance indicator to document window
Fragmented functionality	need ability to access next document from document window instead of having to go back to document list window	H	1 day	Add new buttons and functionality to document window
Unexpected results	when user hits return in keyword box, search is accidentally performed	L	None	No action; leave as is to avoid making user use the mouse to click on “perform search”; address in training

Table 3 - Sample Usability Matrix by Problem Category

Changes were made to the interface and the second usability test was conducted. Figure 3 is a screenshot of the interface after changes were made to address problems identified in the first pass of usability testing. The changes made to the interface are identified and compared to the original interface shown in Figures 1 and 2.

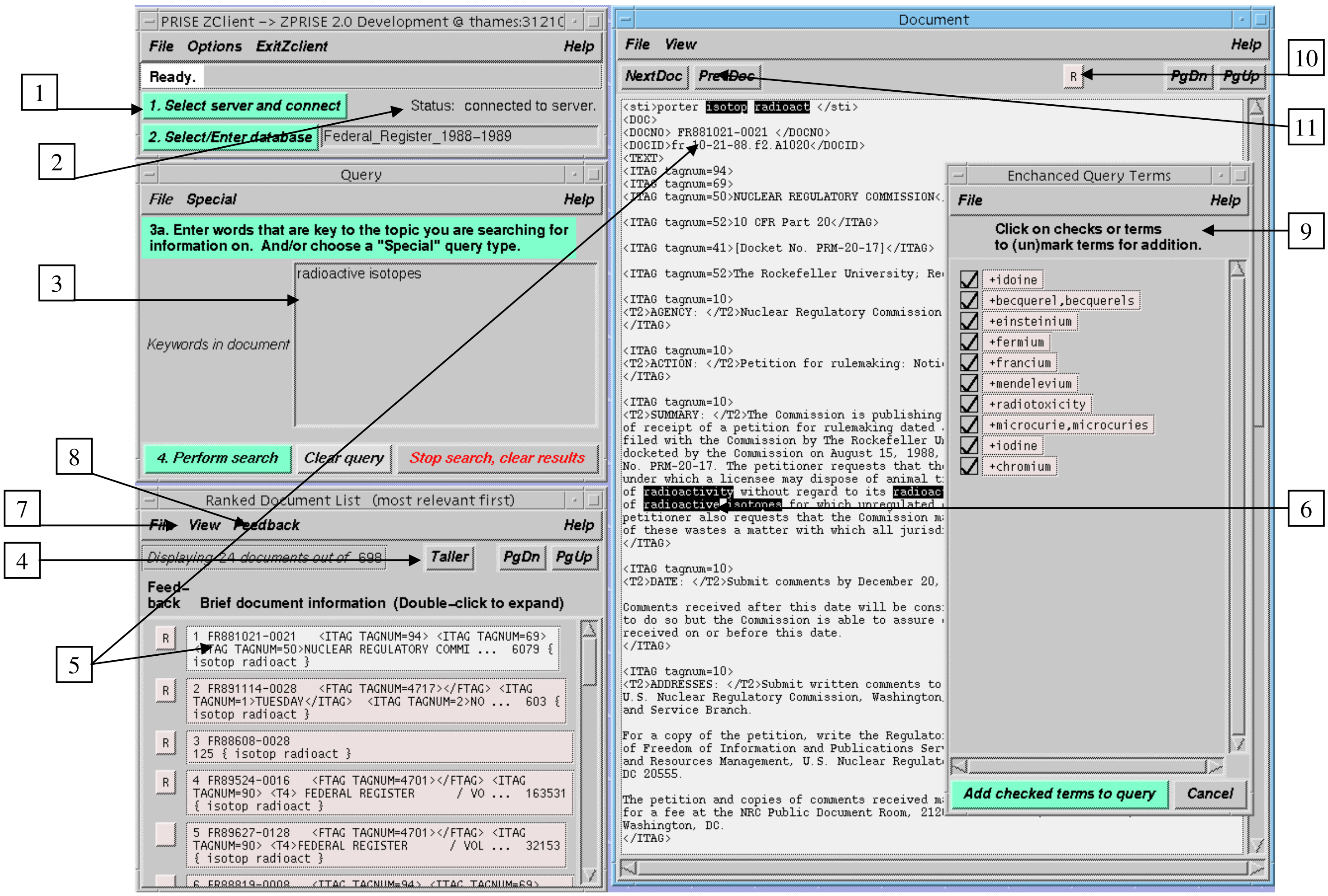

Figure 3 - Interface after changes were made based on results of usability test 1

Based on the results of Test 1, in addition to changing the entire color scheme of the topic development system, several other changes were made to improve the usability of the system. We found that users were often confused about what their first step should be once the topic development system was displayed. To help guide them, we numbered the steps and clarified the instructions. For example, "Server" became "1. Select server and connect" (1). The next change we made was to relocate the status box. Even though the system had a status box which indicated what state the system was in, the users had difficulty seeing the status messages due to the location of the box. Since they were unsure of what the system was doing, they became frustrated and began to press buttons which only made the system take longer to process their requests. Originally the status messages appeared below the database selection button. We moved the status box (2) next to the server selection button and saw improvements in user actions in Test 2.

In the "Query" window, the query box in Test 1 was very small. This gave the users the impression that they could only type in a few query terms even though the system allowed for an unlimited number of terms. In addition, the users wanted to see everything they had typed and the narrow window in the first system did not allow for this. To address this problem, we enlarged the query box (3) which gave the users the opportunity to review the query terms they had entered. This change contributed to increased accuracy in retrieval results since any incorrectly spelled terms were corrected before the search was performed.

In the “Document List” window, users wanted to see the maximum number of document titles available to them. We added a “Taller/Shorter” (4) feature that allowed the user to increase or decrease the number of document titles viewable. When the taller feature is activated, the user can see twice as many document titles as provided by the shorter feature. Making the list taller did not come without a price, however. The “Query” window is covered with the list of document titles when the “Taller” feature is activated. This did not appear to be a deterrent in using the feature.

Another change to the “Document List” window involved the color scheme. The colors used in the system for Test 1 were so similar that it was difficult for the user to tell which document was highlighted in the “Document List” window and displayed in the “Document” window. In the system used for Test 2, the color scheme was changed so that the background color of the document being displayed in the “Document” window now matched the background color of the document title highlighted in the “Ranked Document List” window (5). This made it easier for the user to make the connection between the two screens. In addition, in the “Document” window the query terms were highlighted in Test 1 which made it difficult to quickly and easily identify the terms. This was changed in the system used in Test 2 so that the query terms were displayed in reverse video (6) in the “Document” window.

In Test 1, the system kept a separate record of the relevant documents and displayed them in a pop-up window called “Relevant Document List”, which overlapped the “Document” window. Even though the “Relevant Document List” window could be hidden from view, the users found this feature unacceptable. They wanted to see their relevance judgments and view the document at the same time. Some users wanted to see all of their judgments in one window while other users wanted to see only the relevant judgments. To accommodate the multitude of user preferences, the system used in Test 2 provides the user with 3 options when using the view feature (7) in the “Ranked Document List” window. One option allows the user to view all document titles judged relevant, irrelevant, or unjudged . Another option allows only the relevant and unjudged document titles to appear and the final option allows only the irrelevant and unjudged document titles to be displayed. This gives the user freedom to view their results in an environment that makes sense to them.

In the system used in Test 1, the “Relevant Document List” window contained a “Show enhanced query terms” or feedback feature which displayed the enhanced query terms in a separate window. Our instructions on this window were cryptic and users had no idea what to do with this list of terms. In Test 1, if the user chose to hide the “Relevant Document List“ window, then the feedback feature was hidden too. To correct these problems, the feedback feature (8) was added to the "Ranked Document List" window and the instructions in the "Enhanced Query Terms" window were clarified (9).

Users expressed a strong desire to perform most of their actions in one window and the window of choice was the "Document" window. They wanted to judge the document and go to the next and previous document from this window. The system used in Test 1 required the user to read the document from the "Document" window and then move to the "Document List" window to make a judgment and call up the next document. This was very time consuming for them to move back and forth from the "Document List" window to the "Document" window. The system used in Test 2 streamlined these tasks and allowed the users to view, judge (10), and move to next and previous documents (11) in the "Document" window .

4 Analysis of Usability Test 2

In order to be able to compare and contrast the results from Test 1 and Test 2, we used the same basic analysis technique for Test 2. During comparative analysis, we were concerned with two major questions. Had we minimized/eliminated the problems identified in Test 1? And, did any new or unreported problems occur in Test 2, especially those that may have been introduced due to the changes?

We examined the categories of usability issues resulting from both tests rather than comparing the actual raw numbers to account for the differences in users and the number of users in each test. Usability problems found in Test 1 and Test 2 and their resolutions are listed in Appendix 1.

Appendix 1 illustrates the process used to compare and categorize the usability issues identified in Test 1 and Test 2. Usability issues were identified in 19 categories in Test 1 and 14 categories in Test 2. When we compared the categories we found 11 in common. This translates to the elimination of eight groups of usability issues between Test 1 and Test 2 and the identification of three new groups in Test 2.

We then analyzed the 11 common groups and the three new groups. We found that they could roughly be classified into two major divisions - navigation issues and conceptual issues. Navigation issues included widget co-location, size, placement, and existence. Conceptual issues primarily revolved around the definition and function of relevance feedback including the use and utility of enhanced query terms.

The process of collecting and organizing the data was critical to the success of accomplishing the three high-level goals in this case study. Our goals were to gain experience in conducting usability testing, to examine the usability of the new ZPRISE interface, and to identify problems our users were having with the topic development task. Focusing on each goal and extracting all the needed information for that goal without compromising the objectives of the other two goals proved difficult. During Test 1 we struggled with several classification schemes before deciding on a useful strategy to categorize and analyze the data collected. In the end, the final data organization and analysis was made easier through repeated examination of the data from several different perspectives during Test 1. The perspectives included grouping the tutorial and task observations by user. Then the combined observations were categorized by interface window and then by user.

5 Lessons Learned

Our first goal in conducting the informal usability study was to gain experience in usability testing on information retrieval systems. First and foremost, we learned that performing several activities in tandem can lead to confusion between tasks and more difficult analysis of results. The users were performing actual TREC topic development and in turn we were testing the new interface for general usability while also testing this general use interface on a specific task. Often the lines became blurred.

The second goal of identifying and correcting the problems related to our general use ZPRISE interface was relatively straightforward. During analysis, we identified navigation and conceptual difficulties which were corrected and retested.

The third goal of identifying problems our users were having with the TREC task became the most complex (and interesting) of the three goals. This section will mainly concentrate on lessons learned in that area.

First, as in most usability studies (Koeneman *et al.,* 1995), we identified the typical user issues:

- Users brought their biases and experiences with them.
- They compared the new interface to the old interface, and these comparisons revealed a great deal about an interface's usability. We expect comparisons to continue and view it as a dynamic feedback mechanism for ourselves and the system developers.
- Users responded favorably to hands-on training during the actual task, and this continues to be the best way for users to grasp and retain knowledge of the interface features.
- Improvements in one area may cause unanticipated adverse effects. For example, additional features placed a noticeable burden on the performance of our computer systems, and response time declined.

We also learned a few TREC-specific and task-specific lessons during the usability study:

- Users confused the topic development task and the relevance assessment task because relevance assessment is part of the topic development task, and some of the screens and features of the new interface resembled the old relevance assessment interface.
- Users required a detailed explanation of the objectives of the topic development task before task performance began. In particular, users became anxious when they did not find what they considered to be a good number of relevant documents. Consequently, they may have changed their definition of what constituted a relevant document midstream in their session. This could lead to inconsistencies during the topic development task as well as during the relevance assessment task.

- The second round of usability testing revealed additional relevance assessment system requirements because the users' most recent task had been relevance assessment.

These observations led to three recommendations specific to the TREC task. The first was the requirement for the design and development of two specialized task-specific interfaces (i.e., topic development and relevance assessment) for the TREC assessors. Based on the results of this case study, the ZPRISE interface was modified to more closely match the requirements of the topic development task. Additionally, the relevance assessment system was modified based on the users' comments.

The second specific issue was to effectively convey to the users the importance of consistency in relevance judgments. We revised the training program, stressing to the users the primary goal of providing consistent judgments and discouraging them from stretching their definition of a topic to gain additional relevant documents. While this type of instruction is helpful, it cannot eliminate all relevance judgment errors. Note, however, that these errors do not compromise the quality of the test collections. Relevance judgments are known to vary widely across different people (Schamber, 1994), but experiments have shown that the comparative evaluation of retrieval performance is stable despite substantial differences in relevance judgments (Voorhees, 1998).

And finally it was obvious that the training program also needed to be revised to allow hands-on training for the topic development and relevance assessment interfaces. Most importantly, the revised training program must explain the TREC task in more detail. Once users have a better understanding of the objectives of their two tasks, we believe that this will lead to increased accuracy in relevance assessments. Our training program now includes a discussion session of known issues that the users have brought to our attention in the past. Hands-on training with practical examples has been introduced and well received by the users. We have also developed a visual manual with screenshots to help the users immediately identify the screen they are having difficulty with.

In conclusion, we learned that even an informal usability study produces a substantial amount of useful data. On the basis of this data, we were able to make significant improvements to our general purpose ZPRISE interface. Additionally we were able to identify problems the TREC users were having with the TREC task and gain insights into user preferences in interface design and usability issues.

Acknowledgements--We would like to thank the anonymous referees for their beneficial comments on the initial version of this paper. In addition, we thank our NIST colleagues: Donna Harman and Ellen Voorhees for guidance, helpful comments, and suggestions; Paul Over and Will Rogers for their help in categorizing and resolving user comments; Martin Smith for his help with data collection; Sharon Laskowski and Judy Devaney for their encouragement and support. A special thanks to the NIST assessors who participated in this study, without whose support this work would not have been possible.

References

Harman, D., & Candela, G. (1990). Retrieving Records from a Gigabyte of Text on a Minicomputer Using Statistical Ranking. In *Journal of the American Society for Information Science*. (pp. 582-589).

Hoffman, D. M. & Downey, L. L. (1997). Lessons Learned In An Informal Usability Study. In N. J. Belkin, A. D. Narasimhalu, & P. Willett (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 337). Philadelphia, PA, USA.

Koenemann, Jurgen, Quatrain, Richard, Cool, Colleen, Belkin, Nicholas J. (1995). New Tools and Old Habits: The Interactive Searching Behavior of Expert Online Searchers using INQUERY. In D. K. Harman (Ed.) *Overview of the Third Text Retrieval Conference (TREC-3).* (pp. 145-177). Gaithersburg, MD, USA.

Library of Congress. (1998). *Library of Congress Maintenance Agency page for International Standard Z39.50* [URL]. http://lcweb.loc.gov/z3950/agency/.

Nielsen, Jakob. (1993). *Usability Engineering.* Academic Press, Inc., San Diego,CA, USA.

NIST. (1998a). *Lessons Learned in an Informal Usability Study* [URL]. http://www.itl.nist.gov/div894/894.02/works/papers/sigir-poster.html.

NIST. (1998b). *The Trec Overview Page* [URL]. http://trec.nist.gov/overview.html.

Schamber, Linda (1994). Relevance and information behavior. *Annual Review of Information Science and Technology,* 29, 3-48.

Sparck Jones, K. & van Rijsbergen, C. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.*

Voorhees, E., & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference (TREC-5). In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text Retrieval Conference (TREC-5).* (pp. 1-28). Gaithersburg, MD, USA.

Voorhees, E. M. (1998). Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson and J. Zobel (Eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 315-323). Melbourne, Australia.

Appendix 1 - Usability Problem Categories from Test 1 and Test 2

Usability Problem Categories from Test 1 and Test 2		
Test 1	**Test 2**	**Remarks**
Ambiguous terminology	Ambiguous terminology	only one occurrence found in Test 2 to be fixed in new system
Auto-scrolling	Auto-scrolling	no action taken from Test 1 but other navigation features will be added in future
Check mark		resolved after Test 1 by adding on-screen cues and more help instruction
Desired feature	Desired feature	no action on various individual preferences that were not appropriate for inclusion in future system
Focus problems		no action after Test 1 on general focus model but some features were enlarged to make getting focus easier
Fragmented functionality	Fragmented functionality	partially resolved after Test 1, more features will be added to document window to reduce eye/mouse left-to-right actions
Messages		resolved after Test 1 by using more distinct wording and locating message bar in more prominent place
Miscellaneous confusion	Miscellaneous confusion	general category, no action taken after Test 1, but some of the same issues appeared in Test 2 – interaction order and enhanced query terms
More information	More information	several fixes after Test 1, and only one occurrence during Test 2 that will be fixed in future system

Relevant document window		resolved after Test 1, relevant document window deleted
Relevance feedback confusion	Relevance feedback confusion	a problem with most systems either from general understanding of concept or its implementation in system
Relevance indicator	Relevance indicator	partially addressed after Test 1, mainly revolved around size and location of widget
Restore irrelevant choice		"not relevant choice" restored from user interaction standpoint but it has no meaning to the search engine
Scrolling	Scrolling	partially addressed after Test 1, paging added, scanning will be added to new system
Software bugs		various bugs fixed after Test 1
Speed	Speed	system/network issue but still a user issue
Unexpected results		font issue addressed after Test 1, but accidentally hitting return in keyword box not addressed and did not appear in Test 2
Use of the color green		resolved after Test 1, green removed and color scheme enhanced to provide cues and aesthetics
Visual discrimination	Visual discrimination	partially resolved after Test 1, keywords now displayed in reverse video, Test 2 comment on distinguishing between active windows will be addressed future system
	"hidden" options/parameters	address in training and allow unlimited # of documents to be displayed

	Task issues	a few task issues probably attributed to new users and training
	Enhanced query term confusion	partially attributable to being new to system but users felt terms were not useful and that there were no meanings provided for words they did not know, some similar issues occurred during Test 1 but were classified more under miscellaneous confusion